Frontiers in Genetics

TYPE Original Research
PUBLISHED 18 September 2023
DOI 10.3389/fgene.2023.1271311

Check for updates

OPEN ACCESS

EDITED BY
Rui Yin,
University of Florida, United States

REVIEWED BY
Hailin Chen,
East China Jiaotong University, China
Bo-Wei Zhao,
Chinese Academy of Sciences (CAS),
China

*CORRESPONDENCE
Mengyun Yang,
✉ mengyun_yang@126.com
Jianxin Wang,
✉ jxwang@mail.csu.edu.cn

RECEIVED 02 August 2023
ACCEPTED 23 August 2023
PUBLISHED 18 September 2023

CITATION
Yang M, Yang B, Duan G and Wang J (2023), ITRPCA: a new model for computational drug repositioning based on improved tensor robust principal component analysis.
Front. Genet. 14:1271311.
doi: 10.3389/fgene.2023.1271311

ITRPCA: a new model for computational drug repositioning based on improved tensor robust principal component analysis

Mengyun Yang[1,2]*, Bin Yang[1], Guihua Duan[3] and Jianxin Wang[3]*

[1]School of Mechanical and Energy Engineering, Shaoyang University, Shaoyang, China, [2]School of Computer Science, Hunan First Normal University, Changsha, China, [3]School of Computer Science and Engineering, Central South University, Changsha, China

Background: Drug repositioning is considered a promising drug development strategy with the goal of discovering new uses for existing drugs. Compared with the experimental screening for drug discovery, computational drug repositioning offers lower cost and higher efficiency and, hence, has become a hot issue in bioinformatics. However, there are sparse samples, multi-source information, and even some noises, which makes it difficult to accurately identify potential drug-associated indications.

Methods: In this article, we propose a new scheme with improved tensor robust principal component analysis (ITRPCA) in multi-source data to predict promising drug–disease associations. First, we use a weighted *k*-nearest neighbor (WKNN) approach to increase the overall density of the drug–disease association matrix that will assist in prediction. Second, a drug tensor with five frontal slices and a disease tensor with two frontal slices are constructed using multi-similarity matrices and an updated association matrix. The two target tensors naturally integrate multiple sources of data from the drug-side aspect and the disease-side aspect, respectively. Third, ITRPCA is employed to isolate the low-rank tensor and noise information in the tensor. In this step, an additional range constraint is incorporated to ensure that all the predicted entry values of a low-rank tensor are within the specific interval. Finally, we focus on identifying promising drug indications by analyzing drug–disease association pairs derived from the low-rank drug and low-rank disease tensors.

Results: We evaluate the effectiveness of the ITRPCA method by comparing it with five prominent existing drug repositioning methods. This evaluation is carried out using 10-fold cross-validation and independent testing experiments. Our numerical results show that ITRPCA not only yields higher prediction accuracy but also exhibits remarkable computational efficiency. Furthermore, case studies demonstrate the practical effectiveness of our method.

KEYWORDS

drug repositioning, tensor robust principal component analysis, weighted k-nearest neighbor, low-rank tensor, drug–disease associations

1 Introduction

Over the past few decades, while funding for drug development has seen a substantial surge, the number of newly approved drugs for market release has remained limited. Notably, developing a new drug demands an average of 13.5 years and involves an average expenditure of 1.8 billion (Liu et al., 2020). This process is time-consuming and tremendously expensive and involves high risk (Chong and Sullivan, 2007; Dickson and Gagnon, 2009). Since the approved drugs already possess safety records, tolerance, and pharmacokinetic data of the human body in clinical trials, discovering new clinical indications for commercialized drugs is an important strategy to improve the efficiency of drug development (Ashburn and Thor, 2004). In fact, there have been a few successful repurposed drugs, such as sildenafil, thalidomide, and retinoic acid, which have been widely used in application (Luo et al., 2021).

Using computational methods to discover new uses for established drugs is a crucial aspect of drug repositioning, which is based on the assumption that drugs with similar properties tend to treat similar diseases. With the rapid development of high-throughput technology and continuously generating multi-omics data, there is an increasing focus on crafting computational methods for elevated precision (Wang et al., 2021). These approaches can be classified into four distinct groups: encompassing network-based methods, machine learning-based methods, matrix-based methods, and deep learning-based methods.

Network-based approaches infer the scores of drug–disease pairs by constructing drug and disease heterogeneous biological networks and extracting topological information. The fundamental assumption is guilt by association, whereby if a certain drug can interact with most of the target's neighbors, it is probable that the target will also be able to interact with the same drug and *vice versa*. Based on the guilt-by-association principle, Wang et al. (2013) used *a priori* information about drugs and targets to establish a heterogeneous graph. A heterogeneous graph-based inference (HGBI) model was used to predict new drug–target interactions. Luo et al. (2016) enhanced the quality of the similarity between drugs and diseases by exploiting the existing drug–disease associations. Building on the combined similarity measures, a new bi-random walk algorithm called MBiRW was developed to infer potential associations between drugs and diseases. Qin et al. (2022) proposed a network-based inference model for new emerging diseases, which used genes as a bridge in a tripartite drug–gene–disease network to infer latent drug–disease associations. Additionally, to account for the structures of networks and the biological aspects related to drugs and indications, Zhao et al. (2022) presented a novel graph representation model based on heterogeneous networks, namely, HINGRL. It integrated the biological networks of drugs and diseases to learn the features from both topological and biological perspectives.

Machine learning-based approaches use supervised learning algorithms to identify potential indications for drugs based on input features and known associations (Vamathevan et al., 2019), including logistic regression (Yang et al., 2021), random forests Zhao et al. (2022), and support vector machines (Lavecchia, 2015). Jiang and Huang (2022) proposed a graph representation model based on random forest for drug repositioning. The method identified drug–disease associations by feeding combined features from the molecular association network into a random forest algorithm. Gao et al. (2022) presented a model for predicting associations between drugs and diseases, employing similarity kernel fusion (SKF) to merge diverse similarity kernels for drugs and diseases. This fusion resulted in two integrated similarity kernels, and the scores of association pairs were calculated using the Laplacian regularized least square (LapRLS) algorithm.

Matrix-based methods use the low-rank matrix representation of the drug–disease association space to identify novel associations based on the similarity of their profiles. Yang et al. (2019) developed a bounded nuclear norm regularization (BNNR) method to obtain the low-rank matrix of the drug–disease association. This method efficiently handles noise originating from the drug and disease similarity. Yang et al. (2021) proposed a multi-similarity bilinear matrix factorization (MSBMF) method that dynamically integrated multiple similarities of drug and disease into drug–disease association training. It limited the predicted values of the drug–disease association to non-negative. Huang et al. (2020) proposed a multi-task learning method that used ensemble matrix factorization to predict both treatment associations and non-treatment associations between drug and disease. The proposed method can capture complementary features associated with these two tasks. Yan et al. (2022) proposed a multi-view learning with matrix completion method (MLMC), which is capable of effectively utilizing multi-source similarity matrices. The Laplacian graph regularization was pulled into MLMC to acquire an all-encompassing feature representation derived from the multi-similarity information of drugs and diseases.

Deep learning-based methods typically use neural network models to learn the feature representation of drugs and diseases and use these features to predict new association pairs. Xuan et al. (2019) introduced a bidirectional deep learning model based on the convolutional neural network (CNN) and bi-directional long- and short-term memory (BiLSTM). This framework incorporates both similarities and associations between drugs and diseases in addition to pathways that connect specific drug–disease pairs. This approach effectively integrates raw and topological data between nodes. Combining similarity network fusion (SNF) and neural network (NN) deep learning models, Jarada et al. (2021) proposed a method known as SNF-NN, which was designed to forecast novel drug–disease associations. Yu et al. (2021) proposed a layer attention graph convolutional network model to detect the potential uses of drugs. The model performs graph convolutional processing on a heterogeneous network constructed from drug and disease information, thereby achieving association prediction.

To mine latent association features in multiple similarities and association data, we present an improved tensor robust principal component analysis (ITRPCA) method. First, we integrate the prior information of drug and disease to compute five indicators for drug similarity and two indicators for disease similarity. Considering that validated drug–disease associations are extremely sparse, a weighted *k*-nearest neighbor (WKNN) preprocessing step is employed to enrich the association matrix that aids in prediction. Then, we construct a drug tensor and a disease tensor using multi-similarity matrices and an updated association matrix. Finally, we apply ITRPCA to isolate the low-rank tensor and noise information in these two tensors, respectively. We focus on the drug–disease association pairs in the clean low-rank tensor to infer promising indications for drugs. Figure 1 illustrates the comprehensive workflow of the ITRPCA method. Our method's key contributions are as follows:

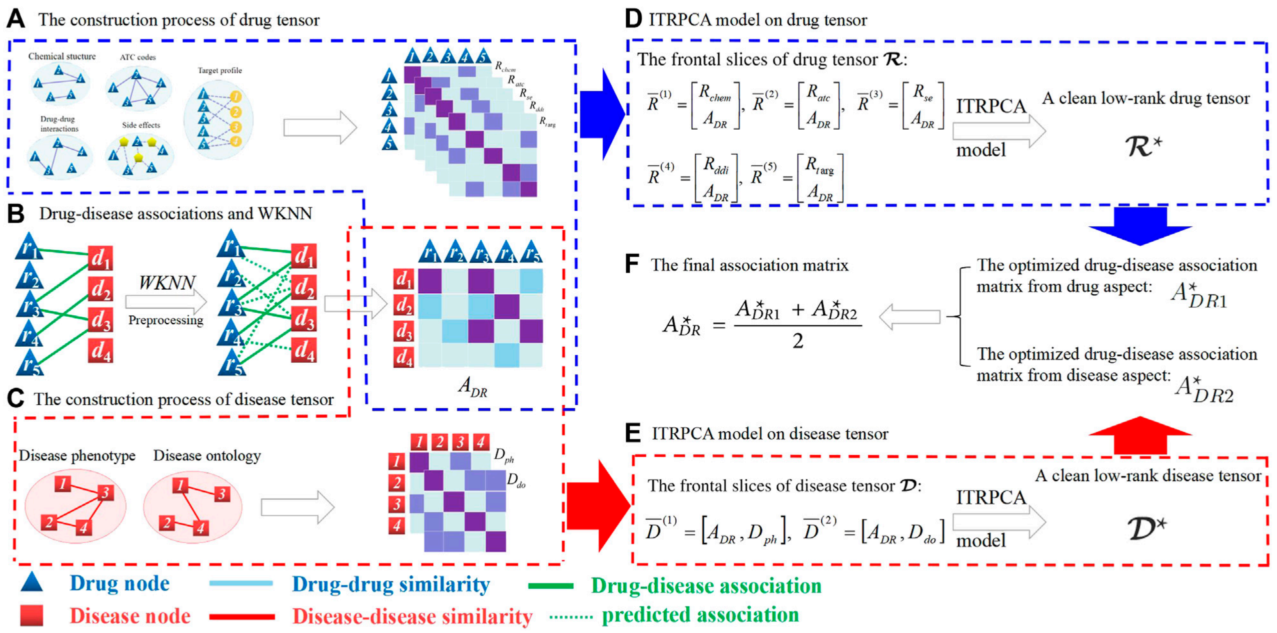
A The construction process of drug tensor
Chemical structure
ATC codes
Target profile
Drug-drug interactions
Side effects
R_{chem} R_{atc} R_{se} R_{ddi} R_{targ}
B Drug-disease associations and WKNN
WKNN
Preprocessing
A_{DR}
C The construction process of disease tensor
Disease phenotype
Disease ontology
D_{ph} D_{do}
D ITRPCA model on drug tensor
The frontal slices of drug tensor $\mathcal{R}$:
$\overline{R}^{(1)}=\begin{bmatrix}R_{chem}\\A_{DR}\end{bmatrix}, \overline{R}^{(2)}=\begin{bmatrix}R_{atc}\\A_{DR}\end{bmatrix}, \overline{R}^{(3)}=\begin{bmatrix}R_{se}\\A_{DR}\end{bmatrix}$
$\overline{R}^{(4)}=\begin{bmatrix}R_{ddi}\\A_{DR}\end{bmatrix}, \overline{R}^{(5)}=\begin{bmatrix}R_{targ}\\A_{DR}\end{bmatrix}$
ITRPCA model
A clean low-rank drug tensor
$\mathcal{R}^*$
F The final association matrix
$A_{DR}^* = \frac{A_{DR1}^* + A_{DR2}^*}{2}$
The optimized drug-disease association matrix from drug aspect: A_{DR1}^*
The optimized drug-disease association matrix from disease aspect: A_{DR2}^*
E ITRPCA model on disease tensor
The frontal slices of disease tensor $\mathcal{D}$:
$\overline{D}^{(1)}=[A_{DR}, D_{ph}], \overline{D}^{(2)}=[A_{DR}, D_{do}]$
ITRPCA model
A clean low-rank disease tensor
$\mathcal{D}^*$
Drug node
Disease node
Drug-drug similarity
Disease-disease similarity
Drug-disease association
predicted association

FIGURE 1
Overall workflow of ITRPCA. **(A)** Construction process of a drug tensor. **(B)** Drug–disease associations and WKNN preprocessing. **(C)** Construction process of a disease tensor. **(D)** ITRPCA model on a drug tensor. **(E)** ITRPCA model on a disease tensor. **(F)** Final association matrix.

- ITRPCA presents a comprehensive scheme for incorporating diverse drug and disease similarities into prediction training.
- By leveraging the weighted tensor Schatten p-norm, ITRPCA can effectively extract the low-rank association tensor from the updated drug and disease tensors, which efficiently separates noisy data and leads to significantly improved accuracy, as demonstrated in our results.
- The ITRPCA model includes a boundary constraint that ensures all predicted tensor entries fall within the predefined interval.
- We have devised an iterative approach employing the augmented Lagrangian multiplier (ALM) to numerically address the ITRPCA model.

2 Materials

To validate the effectiveness of our proposed method, this study involves three crucial datasets: the gold standard dataset (Gottlieb et al., 2011), Cdataset (Luo et al., 2016), and CTD (Davis et al., 2019). Table 1 summarizes the details of these three data, such as the number of drugs and diseases, the number of known association pairs, and the intended purposes in this study. These drugs and diseases are obtained from DrugBank (Wishart et al., 2006) and the Online Mendelian Inheritance in Man (OMIM) database (Ada Hamosh, 2005), respectively. The corresponding drug–disease association matrix denoted as A is represented by a binary matrix, where the proven drug–disease associations are denoted by 1 s, while unproven associations are denoted by 0 s.

Here, we calculate a total of five similarity matrices for drugs: chemical structure similarity R_{chem}, anatomical therapeutic chemical (ATC) code similarity R_{atc}, side-effect similarity R_{se}, drug–drug interactions similarity R_{ddi}, and target profile similarity R_{targ}. Based on the drug's canonical SMILES (Weininger, 1988) file, we use the Chemical Development Kit (CDK) (Steinbeck et al., 2003) tool to compute the hashed fingerprints for all drugs and then obtained R_{chem}. ATC codes for all relevant drugs were extracted from DrugBank. We use the semantic similarity algorithm (Resnik et al., 1995) to calculate the similarity scores between ATC terms and then obtained R_{atc}. The rest of the similarities are calculated using the Jaccard similarity coefficient (Jaccard, 1908), which can be expressed as follows:

TABLE 1 Number of drugs, diseases, and known association pairs in each dataset along with their respective purposes for dataset utilization.

Dataset	# of drugs	# of diseases	# of association pairs	Purpose
Gold standard dataset	593	313	1,933	10-fold cross-validation
Cdataset	663	409	2,352	Independent testing
CTD (February 2020)	1,613	969	15,339	Independent testing

$$R_{se/ddi/targ}(i,j) = \frac{|S_i \cap S_j|}{|S_i \cup S_j|}, \tag{1}$$

where S_i implies the side-effect profiles of drug i in R_{se}, drug–drug interaction profiles of drug i in R_{ddi}, and drug–target interaction profiles of drug i in R_{targ}.

For diseases, a total of two similarity measures are calculated: disease phenotypic similarity D_{ph} and disease ontology similar D_{do}. D_{ph} is obtained from MimMiner (Van Driel et al., 2006), which calculates the frequency of MeSH (medical subject heading (MeSH) vocabulary terms co-occurring in the medical descriptions of two diseases retrieved from the OMIM database. According to the structure of the disease ontology academic language, D_{do} is computed by using the Gene Ontology-based algorithm (Wang et al., 2007).

In summary, we have collected a total of one drug–disease association matrix A, five drug similarity matrices (i.e., R_{chem}, R_{atc}, R_{se}, R_{ddi}, and R_{targ}), and two disease similarity matrices (i.e., D_{ph} and D_{do}) for computational drug repositioning.

3 Methods

In this section, we introduce our method for identifying potential uses for established drugs. The structure is as follows: first, we depict the robust principal component analysis (RPCA) and tensor robust principal component analysis (TRPCA). Then, we propose the model of ITRPCA according to the requirements of drug repositioning. At last, the ALM method is demonstrated to solve the ITRPCA model in detail.

For the ease of reference, bold calligraphy letters represent third-order tensors, e.g., $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$, capital letters denote the matrices, e.g., X, bold lower-case letters indicate the vectors, e.g., $\boldsymbol{x}$, and lower-case $\boldsymbol{\mathcal{X}}_{ijk}$ denote the elements of $\boldsymbol{\mathcal{X}}$.

3.1 Robust principal component analysis

RPCA stands as a prominent technique in low-rank representation, which can separate the noise matrix from the original data matrix and learn the clean low-rank matrix. It has found successful applications in computer vision and machine learning, such as video surveillance (Wright et al., 2009), facial modeling (Peng et al., 2012), and subspace clustering (Liu et al., 2010). RPCA is targeted at a matrix, which can decompose the target matrix into a low-rank matrix and a sparse matrix for achieving noise reduction. Generally, the mathematical formula of RPCA can be expressed as

$$\min_{X,E} \|X\|_* + \lambda\|E\|_1 \quad \text{s.t. } M = X + E, \tag{2}$$

where M denotes the original matrix, X is the low-rank matrix, and E is the sparse noise matrix. $\|X\|_* = \sum_r \sigma_r(X)$ represents the nuclear norm of matrix X, where $\sigma_r(X)$ is the rth singular value of X. $\|E\|_1 = \sum_{ij}|e_{ij}|$ denotes the L_1-norm of E, and e_{ij} is the (i, j) element of E.

3.2 Tensor robust principal component analysis

TRPCA (Lu et al., 2020) is a continuation of RPCA. The primary motivation behind developing TRPCA is to handle multi-dimensional datasets, which are prevalent in various domains, including computer vision (Wang et al., 2014), object recognition (Zhang and Peng, 2019), and medical imaging (Pham et al., 2021). TRPCA aims to decompose the multi-dimensional data into a low-rank tensor, which captures the essential features of the data, and a sparse tensor, which contains the outliers and noise. The low-rank tensor can be interpreted as the underlying structure of the data, while the sparse tensor represents the deviations from this structure.

Similar to the nuclear norm of the matrix, the tensor nuclear norm (Kilmer and Martin, 2011) is defined as

$$\|\boldsymbol{\mathcal{X}}\|_* = \sum_{i=1}^{n_3} \left\|\bar{X}^{(i)}\right\|_* = \sum_{i=1}^{n_3} \sum_{j=1}^{l} \sigma_j\left(\bar{X}^{(i)}\right), \tag{3}$$

where $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ and $l = \min(n_1, n_2)$. $\bar{X}^{(i)}$is denoted as the ith frontal slice of $\boldsymbol{\mathcal{X}}$, and $\bar{\boldsymbol{\mathcal{X}}}$ is denoted as the discrete fast Fourier transform (FFT) of $\boldsymbol{\mathcal{X}}$ along the third dimension, i.e., $\bar{\boldsymbol{\mathcal{X}}} = ifft(\boldsymbol{\mathcal{X}}, [], 3)$. Thus, $\boldsymbol{\mathcal{X}} = ifft(\bar{\boldsymbol{\mathcal{X}}}, [], 3)$. The TRPCA model is formulated as follows:

$$\min_{\boldsymbol{\mathcal{E}},\boldsymbol{\mathcal{X}}} \lambda\|\boldsymbol{\mathcal{E}}\|_1 + \|\boldsymbol{\mathcal{X}}\|_* \quad \text{s.t. } \boldsymbol{\mathcal{M}} = \boldsymbol{\mathcal{X}} + \boldsymbol{\mathcal{E}}, \tag{4}$$

where $\boldsymbol{\mathcal{M}}$ is the original tensor data, $\boldsymbol{\mathcal{X}}$ measures the low-rank tensor, and $\boldsymbol{\mathcal{E}}$ denotes the sparse noise tensor. According to Eq. 3, model (4) is equally regularized for all singular values of the tensor data and shrunk with the same parameters when minimizing the tensor nuclear norm.

3.3 ITRPCA for drug repositioning

Weighted k-nearest neighbor preprocessing. $\{d_1, d_2, \ldots, d_n\}$ and $\{r_1, r_2, \ldots, r_m\}$ represent the collection of n disease nodes and m drug nodes, respectively. $A \in \mathbb{R}^{n \times m}$ represents the original drug–disease association matrix, where $A_{ij} = 1$ if disease d_i is recognized to have a known connection with drug r_j; otherwise, $A_{ij} = 0$. The ith row vector of matrix A, i.e., $A_d(d_i) = (A_{i1}, A_{i2}, \ldots, A_{im})$, represents the association profile of disease d_i. The jth column vector of matrix A, i.e., $A_r(r_j) = (A_{1j}, A_{2j}, \ldots, A_{nj})$, represents the association profile of drug r_j. In fact, if novel drug nodes or disease nodes are considered, the values of their corresponding columns or rows in the adjacency matrix are zero. This case will lead to unsatisfactory performance in prediction (Xiao et al., 2018). We utilize the WKNN algorithm to populate the drug–disease association matrix. This is achieved by considering the similarities of drugs and diseases.

For each drug r_q, the similarities of the other k-nearest known drugs (where at least one validated association exists) are combined to update the drug's association profile:

$$A_r(r_q) = \frac{1}{Q_r} \sum_{j=1}^{K} \alpha^{j-1} R(r_j, r_q) A_r(r_j), \tag{5}$$

where the drugs r_1 to r_k are arranged in descending order according to their similarity with r_q. $\alpha \in [0, 1]$ is a decay term, and R denotes the mean matrix of five drug similarity matrices. This means that when the similarity between r_j and r_q is strong, a higher weight will be assigned; conversely, a lower weight will be assigned. Furthermore, $Q_r = \sum_{1 \le j \le k} R(r_j, r_q)$ is the normalization term.

In the same way, the updated association profile for each disease d_p is obtained as follows:

$$A_d\left(d_p\right) = \frac{1}{Q_d}\sum_{i=1}^{K}\alpha^{i-1}D\left(d_i, d_p\right)A_d\left(d_i\right), \tag{6}$$

where d_1 to d_k are the diseases sorted in descending order based on their similarity to d_p. $\alpha \in [0, 1]$ is a decay term, and D denotes the mean matrix of two disease similarity matrices. Q_d is a normalization term, and $Q_d = \sum_{1\le i\le k} D(d_i, d_p)$. Finishing these profile operations, we obtain the aforementioned two matrices A_r and A_d from drug and disease spaces, respectively. Then, the new drug–disease association matrix A_{DR} is calculated as follows:

$$A_{DR} = \max\left(A, \frac{A_r + A_d}{2}\right). \tag{7}$$

After the processing of WKNN, the density of the updated association matrix A_{DR} is greatly improved, and it no longer contains all zero rows and all zero columns. However, some noise information is inevitably added into the association matrix. Subsequently, we will propose our new method for noise separation.

Algorithm 1 summarizes the preprocessing step for updating the drug–disease association matrix using WKNN.

```
• Input: The original drug–disease association matrix $A \in \mathbb{R}^{n\times m}$, the five drug similarity matrices: $R_{chem}$, $R_{atc}$, $R_{se}$, $R_{ddi}$, $R_{targ}$; the two disease similarity matrices: $D_{ph}$, $D_{do}$, decay term $\alpha$, neighborhood sizes $k$.
• Output: Optimized association matrix $A_{DR}$.
$R = (R_{chem} + R_{atc} + R_{se} + R_{ddi} + R_{targ})/5$, $D = (D_{ph} + D_{do})/2$
for each drug $r_q$ do
    $V = KNN(r_q, k, R)$; //$KNN(r_q, k, R)$ is the function to obtain the $k$ known nearest neighbors of $r_q$ in matrix $R$ in descending order.
    $Q_r = \sum_{j=1}^{k} R(r_j, r_q)$;
    $A_r(r_q) = \sum_{j=1}^{k}\alpha^{j-1}R(r_j, r_q)A(r_j)/Q_r$; //$r_j \in V$
end for
for each disease $d_p$ do
    $U = KNN(d_p, k, D)$;
    $Q_d = \sum_{i=1}^{k} D(d_i, d_p)$;
    $A_d(d_p) = \sum_{i=1}^{k}\alpha^{i-1}D(d_i, d_p)A(d_i)/Q_d$; //$d_i \in U$
end for
$A_{DR} = \max(A, \frac{A_r+A_d}{2})$;
return $A_{DR}$.
```

Algorithm 1 : WKNN preprocessing step for updating the association matrix.

Drug tensor and disease tensor. We construct a third-order drug tensor with five frontal slices denoted as $\boldsymbol{\mathcal{R}} \in \mathbb{R}^{(m+n)\times m\times 5}$. This tensor comprised five drug similarity matrices and an updated association matrix. Specifically, the first frontal slice of the drug tensor is a concatenation of R_{chem} and A_{DR}, which can be described as follows:

$$\bar{R}^{(1)} = \begin{bmatrix} R_{chem} \\ A_{DR} \end{bmatrix}, \tag{8}$$

where $\bar{R}^{(1)} \in \mathbb{R}^{(m+n)\times m}$. In the same way, the remaining four frontal slices of the drug tensor can be constructed with other similarity matrices and A_{DR}, which is presented as

$$\begin{gathered} \bar{R}^{(2)} = \begin{bmatrix} R_{atc} \\ A_{DR} \end{bmatrix}, \bar{R}^{(3)} = \begin{bmatrix} R_{se} \\ A_{DR} \end{bmatrix}, \\ \bar{R}^{(4)} = \begin{bmatrix} R_{ddi} \\ A_{DR} \end{bmatrix}, \bar{R}^{(5)} = \begin{bmatrix} R_{targ} \\ A_{DR} \end{bmatrix}. \end{gathered} \tag{9}$$

A third-order disease tensor with two frontal slices, namely, $\boldsymbol{\mathcal{D}} \in \mathbb{R}^{n\times (m+n)\times 2}$, is constructed using two disease similarity matrices and an updated association matrix. The disease tensor $\boldsymbol{\mathcal{D}}$ is stacked by two slices. Each of its slices can be denoted as

$$\begin{gathered} \bar{D}^{(1)} = \begin{bmatrix} A_{DR} & D_{ph} \end{bmatrix}, \\ \bar{D}^{(2)} = \begin{bmatrix} A_{DR} & D_{do} \end{bmatrix}, \end{gathered} \tag{10}$$

where $\bar{D}^{(1)}, \bar{D}^{(2)} \in \mathbb{R}^{n\times (m+n)}$.

ITRPCA model. In the two tensors, $\boldsymbol{\mathcal{R}}$ and $\boldsymbol{\mathcal{D}}$, some noise is involved in both the similarity data and the inferred association data by WKNN. TRPCA can be employed to separate noise tensors from low-rank tensors. In order to fully exploit the significant information embedded within drug and disease tensors, it is crucial that we adjust the shrinking of large and small singular values such that the large singular values shrink less and the small singular values shrink more. However, TRPCA fails to effectively utilize this prior knowledge during the minimization of tensor nuclear norm. Therefore, the weighted tensor Schatten p-norm is introduced to treat different singular values separately, which is defined as

$$\begin{aligned} \|\boldsymbol{\mathcal{X}}\|_{\boldsymbol{\omega},S_p} &= \left(\sum_{i=1}^{n_3}\left\|\bar{X}^{(i)}\right\|_{\boldsymbol{\omega},S_p}^{p}\right)^{\frac{1}{p}} \\ &= \left(\sum_{i=1}^{n_3}\sum_{j=1}^{h}\omega_j * \sigma_j\left(\bar{X}^{(i)}\right)^{p}\right)^{\frac{1}{p}}, \end{aligned} \tag{11}$$

where $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{n_1\times n_2\times n_3}$, $h = \min(n_1, n_2)$, σ_j denotes the jth singular value, and ω_j denotes the weight value of the jth singular value. When p = 1 and $\boldsymbol{\omega} = 1$, $\|\boldsymbol{\mathcal{X}}\|_*$ is a special case of $\|\boldsymbol{\mathcal{X}}\|_{\boldsymbol{\omega},S_p}$. Moreover, it is crucial to note that the entries of the low-rank tensor using TRPCA can be any real value in the range of $(-\infty, +\infty)$. However, it is imperative to ensure that the predicted values are contextually relevant, as any values falling outside the interval of [0,1] would be meaningless. To address this concern, a bound constraint should be incorporated to restrict the predicted values of unobserved elements within the interval [0, 1]. Our ITRPCA model is formulated as follows:

$$\begin{gathered} \min_{\boldsymbol{\mathcal{E}},\boldsymbol{\mathcal{X}}} \lambda\|\boldsymbol{\mathcal{E}}\|_1 + \|\boldsymbol{\mathcal{X}}\|_{\boldsymbol{\omega},S_p}^{p} \\ s.t.\ \boldsymbol{\mathcal{M}} = \boldsymbol{\mathcal{X}} + \boldsymbol{\mathcal{E}} \\ 0 \le \boldsymbol{\mathcal{X}} \le 1, \end{gathered} \tag{12}$$

where $\boldsymbol{\mathcal{M}}$ can be replaced by a drug tensor $\boldsymbol{\mathcal{R}}$ and disease tensor $\boldsymbol{\mathcal{D}}$ in practice.

Here, we use the drug tensor $\boldsymbol{\mathcal{R}}$ instead of $\boldsymbol{\mathcal{M}}$ as an example. By optimizing the ITRPCA model, a clean low-rank drug tensor $\boldsymbol{\mathcal{R}}^* \in \mathbb{R}^{(m+n)\times m\times 5}$ can be obtained. Its potential low-rank representation comes from drug multiple similarity data and association information. Actually, we focus on the part of the association tensor in $\boldsymbol{\mathcal{R}}^*$, which is denoted as $\boldsymbol{\mathcal{A}}_{DR1}^{\star}$ and equal to $\boldsymbol{\mathcal{R}}^*(m+1\colon n+m, :, :)$. In order to obtain a predicted association matrix for inferring potential drug–disease pairs, we take the average of the tensor $\boldsymbol{\mathcal{A}}_{DR1}^{\star}$ in the longitudinal direction. This operation can be expressed as $A_{DR1}^{\star} = avg(\boldsymbol{\mathcal{A}}_{DR1}^{\star}, 3)$, where $A_{DR1}^{\star}$ is the optimized drug–disease association matrix from the drug's perspective. In the same manner, we substitute the disease tensor $\mathcal{D}$ for $\mathcal{M}$ in model (12). A new low-rank tensor $\mathcal{D}^{*}$, the other association tensor $\mathcal{A}_{DR2}^{*}$, and the corresponding association matrix A_{DR2}^{*} can be conducted from the perspective of diseases using ITRPCA. It should be noted that $\mathcal{A}_{DR2}^{*} = \mathcal{D}^{*}(:,1:m,:)$ and $A_{DR2}^{*} = avg(\mathcal{A}_{DR2}^{*},3)$. Finally, the integrated drug–disease association matrix A_{DR}^{*} was obtained by averaging the prediction results of both drugs and diseases.

$$A_{DR}^{*} = \frac{A_{DR1}^{*} + A_{DR2}^{*}}{2}. \tag{13}$$

Algorithm 2 summarizes the process of applying ITRPCA in drug repositioning. Based on the predicted pair scores in A_{DR}^{*}, the potential drug–disease association can be inferred.

```
• Input: Original drug–disease association matrix
  $A \in \mathbb{R}^{n\times m}$, the mean of drug multiple similarity
  $R \in \mathbb{R}^{m\times m}$, the mean of disease multiple similarity
  $D \in \mathbb{R}^{n\times n}$, neighborhood sizes $k$, p-value of Schatten
  p-norm.
• Output: Predicted association matrix $A_{DR}^{*}$.
1. $A_{DR} \leftarrow$WKNN preprocessing $(A, R, k)$;
2. Assign $\mathcal{R}$ and $\mathcal{D}$ by the Eqs (8 - 10).
3. $\mathcal{R}^{*} \leftarrow$ ITRPCA$(\mathcal{R}, p)$;
4. $\mathcal{A}_{DR1}^{*} = \mathcal{R}^{*}(m+1:n+m,:,:)$;
5. $\mathcal{D}^{*} \leftarrow ITRPCA(\mathcal{D}, p)$;
6. $\mathcal{A}_{DR2}^{*} = \mathcal{D}^{*}(:,1:m,:)$;
7. $A_{DR1}^{*} \leftarrow avg(\mathcal{A}_{DR1}^{*},3), A_{DR2}^{*} \leftarrow avg(\mathcal{A}_{DR2}^{*},3)$;
8. $A_{DR}^{*} = \frac{A_{DR1}^{*} + A_{DR2}^{*}}{2}$;
9. return $A_{DR}^{*}$.
```

Algorithm 2 : ITRPCA algorithm in drug repositioning.

3.4 Solutions for ITRPCA

In this subsection, the ALM method is derived to solve the model (12). Accordingly, the augmented Lagrangian function becomes

$$\begin{aligned}\Gamma(\mathcal{E},\mathcal{X},\mathcal{L},\mu) = \lambda\|\mathcal{E}\|_1 + \langle\mathcal{L},\mathcal{M}-\mathcal{X}-\mathcal{E}\rangle \\ +\|\mathcal{X}\|_{\omega,S_p}^{p} + \frac{\mu}{2}\|\mathcal{M}-\mathcal{X}-\mathcal{E}\|_F^2,\end{aligned} \tag{14}$$

where $\mathcal{L}$ is the Lagrange multiplier and μ is the penalty parameter. The primary procedure comprises the subsequent distinct subtasks:

Compute $\mathcal{E}_{k+1}$:We fix $\mathcal{X}_k$ and $\mathcal{L}_k$ to minimize $\Gamma(\mathcal{E},\mathcal{X}_k,\mathcal{L}_k,\mu_k)$ for $\mathcal{E}_{k+1}$. The model (14) becomes

$$\underset{\mathcal{E}}{\arg\min}\ \frac{\lambda}{\mu_k}\|\mathcal{E}\|_1 + \frac{1}{2}\|\mathcal{E}-\boldsymbol{H}_k\|_F^2, \tag{15}$$

where $\mathcal{H}_k = \mathcal{M} + \mu_k^{-1}\mathcal{L}_k - \mathcal{X}_k$, and drawing inspiration from the soft-thresholding operator, we have

$$\mathcal{E}_{k+1} = \mathrm{T}_{\frac{\lambda}{\mu_k}}(\mathcal{H}_k), \tag{16}$$

where the (i, j, k)th element of $\mathrm{T}_{\frac{\lambda}{\mu_k}}(\mathcal{H}_k)$ is $\mathrm{sign}((\mathcal{H}_k)_{i,j,k})\bullet\ \max(|(\mathcal{H}_k)_{i,j,k}| - \lambda/\mu_k, 0)$.

TABLE 2 Sum of AUC and AUPR values using different *k*- and *p*-values in the 10-fold cross-validation.

p \ *k*	10	20	30	40	50
0.6	1.294	1.299	1.302	1.308	1.312
0.7	1.318	1.327	1.335	1.342	1.348
0.8	1.352	1.365	1.370	1.377	1.378
0.9	1.379	1.391	**1.395**	1.394	1.394
1	1.359	1.364	1.366	1.366	1.365

Compute $\mathcal{X}_{k+1}$: We fix $\mathcal{E}_{k+1}$ and $\mathcal{L}_k$ to minimize $\Gamma(\mathcal{E}_{k+1},\mathcal{X},\mathcal{L}_k,\mu_k)$ for $\mathcal{X}_{k+1}$. The model (14) becomes

$$\underset{\mathcal{X}}{\arg\min}\ \mu_k^{-1}\|\mathcal{X}\|_{\omega,S_p}^{p} + \frac{1}{2}\|\mathcal{X}-\mathcal{Y}_k\|_F^2, \tag{17}$$

where $\mathcal{Y}_k = \mathcal{M} + \mu_k^{-1}\mathcal{L}_k - \mathcal{E}_{k+1}$. This is a weighted tensor Schatten p-norm minimization (WTSNM) problem based on t-SVD (Kilmer and Martin, 2011). In order to tackle this concern, the subsequent lemma and theorems can be employed.

Lemma 1 (Xie et al., 2016). For the optimization problem

$$\min_{\delta\geq 0} f(\delta) = \frac{1}{2}(\delta-\sigma)^2 + \omega\delta^p, \tag{18}$$

with the given p and ω, there exists a specific threshold

$$\tau_p^{GST}(\omega) = (2\omega(1-p))^{\frac{1}{2-p}} + \omega p(2\omega(1-p))^{\frac{p-1}{2-p}}, \tag{19}$$

and we have the following conclusions:

1) When $\sigma \leq \tau_p^{GST}(\omega)$, the optimal solution $T_p^{GST}(\sigma,\omega)$ of Eq. 18 is 0.
2) When $\sigma > \tau_p^{GST}(\omega)$, the optimal solution is $T_p^{GST}(\sigma,\omega) = \mathrm{sign}(\sigma)S_p^{GST}(\sigma,\omega)$, and $S_p^{GST}(\sigma,\omega)$ can be obtain by solving $S_p^{GST}(\sigma,\omega) - \sigma + \omega p(S_p^{GST}(\sigma,\omega))^{p-1} = 0$.

Theorem 1 (Xie et al., 2016). Let $Y = U_Y D_Y V_Y^T$ be the SVD of $Y \in \mathbb{R}^{m\times n}$, $\tau > 0$, $l = \min(m, n)$, $0 \leq \omega_1 \leq \omega_2 \leq \cdots \leq \omega_l$, then a global optimal solution of the following model:

$$\underset{X}{\arg\min}\ \frac{1}{2}\|X-Y\|_F^2 + \tau\|X\|_{\omega,S_p}^{p}, \tag{20}$$

is

$$\Upsilon_{\tau*\omega}[Y] = U_Y P_{\tau*\omega}(Y) V_Y^T, \tag{21}$$

where $P_{\tau*\omega}(Y) = \mathrm{diag}(\gamma_1,\gamma_2,\ldots,\gamma_l)$ and $\gamma_i = T_p^{GST}(\sigma_i(Y),\tau*\omega_i)$, which can be obtained by Lemma 1. The $\{\sigma_i(Y)\}$ is organized in a descending order, while $\{\omega_i\}$ is arranged in an ascending order.

Theorem 2 (Gao et al., 2021). Suppose $\mathcal{A} \in \mathbb{R}^{n_1\times n_2\times n_3}$, $l = \min(n_1, n_2)$, $0 \leq \omega_1 \leq \omega_2 \leq \cdots \leq \omega_l$, let $\mathcal{A} = \mathcal{U}*\mathcal{S}*\mathcal{V}^T$ given the model

$$\underset{\mathcal{X}}{\arg\min}\ \frac{1}{2}\|\mathcal{X}-\mathcal{A}\|_F^2 + \tau\|\mathcal{X}\|_{\omega,S_p}^{p}. \tag{22}$$

Then, a global optimal solution to the model (22) is

$$\mathcal{X}^{*} = \Upsilon_{\tau*\omega}(\mathcal{A}) = \mathcal{U}*ifft\left(P_{\tau*\omega}\left(\bar{\mathcal{A}}\right)\right)*\mathcal{V}^T, \tag{23}$$

where $P_{\tau*\omega}(\bar{\mathcal{A}})$ is a tensor and $P_{\tau*\omega}(\bar{A}^{(i)})$ is the ith frontal slice of $P_{\tau*\omega}(\bar{\mathcal{A}})$. $\mathcal{U} = ifft(\bar{\mathcal{U}}, [], 3)$ and $\mathcal{V} = ifft(\bar{\mathcal{V}}, [], 3)$.

TABLE 3 AUC, AUPR, and precision values of all comparison methods in 10-fold cross-validation for the gold standard dataset.

Metric	ITRPCA	HGBI	MBiRW	BNNR	MSBMF	MLMC
AUC	**0.952**	0.829	0.917	0.932	0.941	0.951
AUPR	**0.442**	0.102	0.264	0.423	0.421	0.436
Precision	**0.476**	0.130	0.304	0.463	0.455	0.475

The most optimal outcomes are indicated in **bold**, while the second-best results are underlined.

According to Theorem 2, the global optimal solution of model (17) is

$$\mathcal{X}^{*} = \Upsilon_{\mu_k^{-1}*\omega}(\mathcal{Y}_k) = \mathcal{U}*ifft\left(P_{\mu_k^{-1}*\omega}\left(\overline{\mathcal{Y}}_k\right)\right)*\mathcal{V}^{T}. \tag{24}$$

In addition, we limit the entry values of $\mathcal{X}_{k+1}$ to the interval [0,1] by using the following projection operator.

$$\boldsymbol{\mathcal{X}_{k+1}} = \mathcal{Q}_{[0,1]}(\mathcal{X}^{*}), \tag{25}$$

where $\mathcal{Q}_{[0,1]}$ is defined as

$$\left(\mathcal{Q}_{[0,1]}(\mathcal{X}^{*})\right)_{ijk} = \begin{cases} 1, & \mathcal{X}^{*}_{ijk} > 1, \\ \mathcal{X}^{*}_{ijk}, & 0 \le \mathcal{X}^{*}_{ijk} \le 1, \\ 0, & \mathcal{X}^{*}_{ijk} < 0. \end{cases} \tag{26}$$

Compute $\mathcal{L}_{k+1}$: We fix $\mathcal{E}_{k+1}$ and $\mathcal{X}_{k+1}$ to minimize $\Gamma(\boldsymbol{\mathcal{E}_{k+1}}, \boldsymbol{\mathcal{X}_{k+1}}, \mathcal{L}, \mu)$ for $\mathcal{L}_{k+1}$. The model (14) becomes

$$\mathcal{L}_{k+1} = \mathcal{L}_k + \mu_k(\mathcal{M} - \boldsymbol{X}_{k+1} - \mathcal{E}_{k+1}). \tag{27}$$

Compute $\boldsymbol{\mu}_{k+1}$: In the ITRPCA model, we employ a scheme that gradually increases the learning rate to facilitate fast convergence (Gao et al., 2021). The penalty parameter becomes

$$\boldsymbol{\mu}_{k+1} = \min(\rho\mu_k, \mu_{\max}). \tag{28}$$

Algorithm 3 provides the overall iterative scheme of the ITRPCA model. It can extract significant information from the drug tensor and disease tensor and ensure that the predicted drug–disease association values are within [0,1].

```
    • Input: Tensor data $\mathcal{M}$ (using drug tensor $\mathcal{R}$ or
      disease tensor $\mathcal{D}$), $p$-value of Schatten p-norm.
    • Output: Low-rank tensor $\mathcal{X}$.
    • Initialize: $\mathcal{X}_0 = \mathcal{E}_0 = \mathcal{L}_0 = \mathbf{0}, \mu_0 = 1e-4, \mu_{\max} = 1e10$,
      $\rho = 1.1$, regularization coefficient $\lambda$ and weight
      vector $\boldsymbol{\omega}$.
repeat
1. Update $\mathcal{E}_{k+1}$ by Eq. 16.
2. Update $\mathcal{X}^{*}$ by Eq. 24.
3. $\mathcal{X}_{k+1} \leftarrow \mathcal{Q}_{[0,1]}(\mathcal{X}^{*})$.
4. Update $\mathcal{L}_{k+1}$ by Eq. 27.
5. Update $\mu_{k+1}$ by Eq. 28.
6. $k \leftarrow k + 1$.
until convergence
returnX.
```

Algorithm 3 Solution for the ITRPCA model.

4 Results and discussion

4.1 Evaluation metrics

To evaluate the effectiveness of ITRPCA, we employ 10-fold cross-validation to predict potential indications for existing drugs. In this process, known drug–disease associations within the gold standard

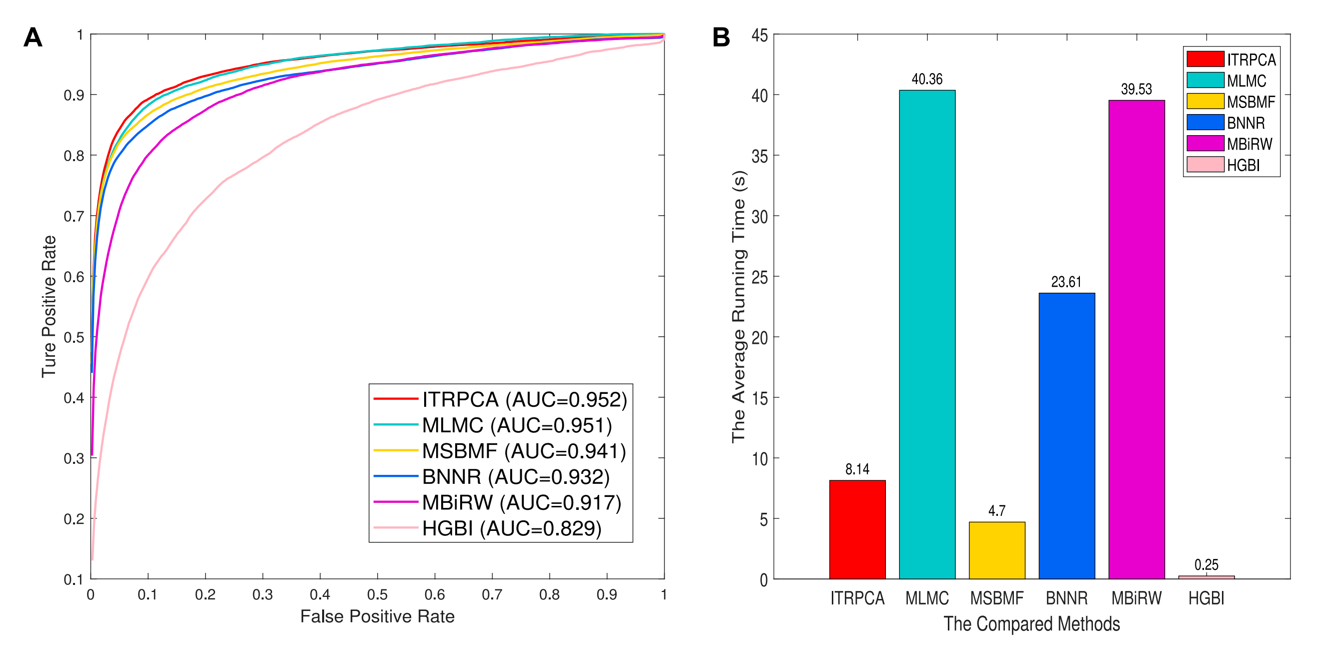

FIGURE 2
Prediction results of all methods for 10-fold cross-validation on the gold standard dataset. **(A)** Receiver operating characteristic curve of prediction results. **(B)** Average running time for each of the 10 folds.

TABLE 4 ***p*-values obtained through Wilcoxon rank sum tests and Bonferroni correction, comparing ITRPCA with other methods on AUC, AUPR, and precision.**

p-value	MLMC	MSBMF	BNNR	MBiRW	HGBI
AUC	1.899	1.125*e*-09	3.019*e*-22	4.932*e*-32	1.281*e*-33
AUPR	1.154	4.741*e*-05	4.741*e*-05	4.741*e*-05	4.741*e*-05
Precision	3.846	4.554*e*-05	1.433*e*-11	1.235*e*-33	1.227*e*-33

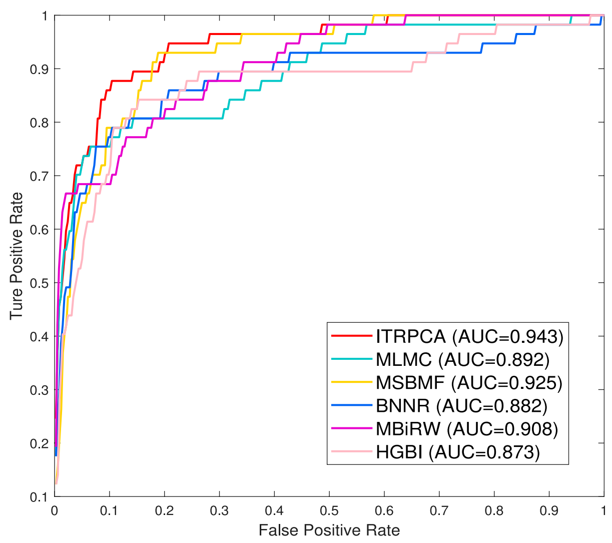

FIGURE 3
ROC curves for all comparison methods tested independently on the Cdataset.

dataset are randomly split into 10 distinct sets of comparable sizes. One subset serves as the test data, while the remaining nine subsets serve as the training data. This 10-fold cross-validation is repeated 10 times with varied random splits, and the resultant averages are considered the final results. Following prediction generation, potential diseases associated with the test drug are sorted in descending order according to their prediction scores. We utilize three evaluation metrics to evaluate the overall performance of ITRPCA: the area under the receiver operating characteristic curve (AUC), the area under the precision–recall curve (AUPR), and precision.

4.2 Parameter setting

In the ITRPCA algorithm, there are some default parameters and two key hyperparameters that need to be adjusted. These default parameters are determined empirically. Specifically, in the WKNN step, we set a decay term α equal to 0.95. In model (12), the regularization coefficient $\lambda = \frac{1}{5\sqrt{n_1 n_2 n_3}}$, where n_1, n_2, and n_3 are the size of $\boldsymbol{\mathcal{X}}$. For drug and disease tensors, we design an adaptive scheme to determine the weight vector $\boldsymbol{\omega}$ of model (12). We divide $\boldsymbol{\omega}$ into three parts: the first part ranges from 1 to u, the second part ranges from $u + 1$ to v, the third part ranges from $v + 1$ to the end, and the specific weights of each part are [1, 2, 4]. The number of u and v are determined by

$$U = avg\left(u_1, u_2, \ldots, u_{n_3}\right), V = avg\left(v_1, v_2, \ldots, v_{n_3}\right), \tag{29}$$

where

$$u_j = \arg\min_x \left\{ \frac{\sum_{i=1}^{x} \sigma_{j,i}}{\sum_{i=1}^{m} \sigma_{j,i}} \geq 0.1 \right\}, \quad j = 1, 2, \ldots, n_3, \tag{30}$$

$$v_j = \arg\min_x \left\{ \frac{\sum_{i=1}^{x} \sigma_{j,i}}{\sum_{i=1}^{m} \sigma_{j,i}} \geq 0.2 \right\}, \quad j = 1, 2, \ldots, n_3. \tag{31}$$

Actually, $\sigma_{j,i}$ is the ith largest singular value of the jth frontal slice matrix of $\boldsymbol{\mathcal{X}}$. It is evident that by minimizing the weighted tensor Schatten p-norm, the singular values of the second and third parts can be shrunk more compared to the first part. The reason is that these two parts are assigned weight values greater than 1. In addition, the two key hyperparameters are needed to be adjusted, which are neighborhood sizes k and p value of Schatten p-norm. We perform grid search to select the appropriate values according to the sum of AUC and AUPR in cross-validation. k is chosen from {10, 20, 30, 40, 50}, and p is picked from {0.6, 0.7, 0.8, 0.9, 1}. The numerical results for determining the parameters k and p are reported in Table 2. When k = 30 and p = 0.9, the highest rating value appears. Meanwhile, we terminate the ITRPCA algorithm when the following stopping criterions are satisfied or the maximum number of iteration steps is reached.

$$f_k \leq tol1, \frac{|f_{k+1} - f_k|}{max\{1, |f_k|\}} \leq tol2, \tag{32}$$

where $f_k = \frac{\|\boldsymbol{\mathcal{X}}_{k+1} - \boldsymbol{\mathcal{X}}_k\|_F}{\|\boldsymbol{\mathcal{X}}_k\|_F}$ and $tol1$ and $tol2$ are the given tolerances, which are set as 10^{-3} and 10^{-4} in the algorithm, respectively.

4.3 Comparison with state-of-the-art drug repositioning methods

We compare ITRPCA with five state-of-the-art methods in computational drug repositioning: HGBI (Wang et al., 2013), MBiRW (Luo et al., 2016), BNNR (Yang et al., 2019), MSBMF (Yang et al., 2021), and MLMC (Yan et al., 2022). To ensure a fair comparison, the parameters used in these compared methods are set to the recommended values by the authors (HGBI: $\alpha = 0.4$; MBiRW: $\alpha = 0.3$ and $l = r = 2$; and BNNR: $\alpha = 1$ and $\beta = 10$) or determined by a grid search (MSBMF: λ_1 and λ_2 are chosen from {0.001, 0.01, 0.1, 1}, and $\tau = 0.7$; MLMC: λ_r and λ_d are selected from {0.0001, 0.001, 0.01, 0.1, 1}, and threshold = 0.8).

We assess the performance of all methods in a 10-fold cross-validation for the gold standard dataset. Table 3 shows the AUC, AUPR, and precision values of all compared methods. As shown in Table 3, ITRPCA has the best performance compared to other methods in terms of AUC, AUPR, and precision. Specifically, ITRPCA achieves the best AUPR value of 0.442, which is 67.424%, 4.492%, 4.988%, and 1.376% higher than the corresponding AUPRs of MBiRW, BNNR, MSBMF, and MLMC, respectively. It can be seen that ITRPCA performs slightly better than MLMC. The ROC curves of all methods in the 10-fold cross-validation are shown in Figure 2A.

Based on the test results from the repeated 10-fold cross-validation, we used the Wilcoxon rank sum tests to evaluate the

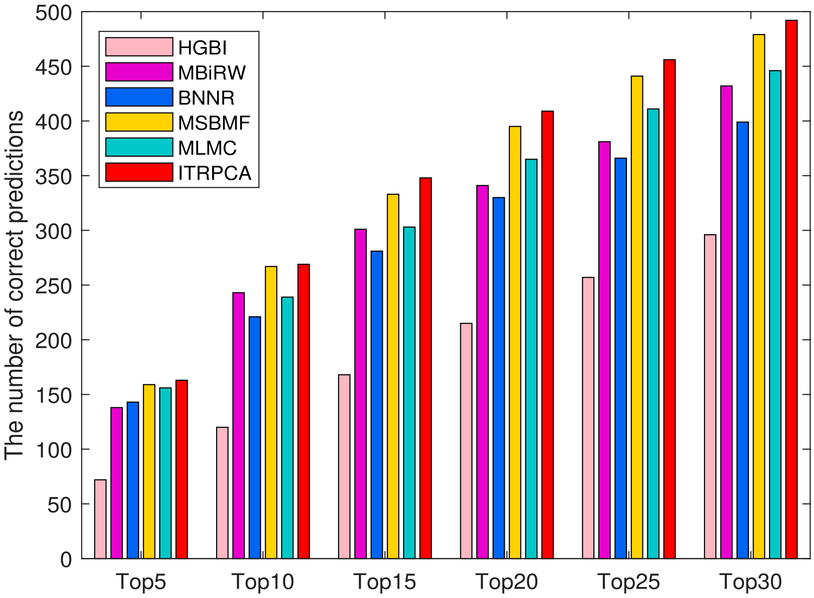

FIGURE 4
Number of top 5 to top 30 indications correctly predicted for all drugs by all comparison methods in the CTD. The *x*-axis represents the comparison of different methods across six specific top n scenarios. The *y*-axis represents the cumulative sum of confirmed indications among the top n predicted indications for each drug, as determined by the respective methods.

TABLE 5 Compared results of ITRPCA, "w/o WKNN," "only WKNN," "ITRPCA-drug," and "ITRPCA-disease" with 10-fold cross-validation on the gold standard dataset.

Metric	ITRPCA	w/o WKNN	only WKNN	ITRPCA-drug	ITRPCA-disease
AUC	**0.952**	0.929	0.903	0.940	<u>0.950</u>
AUPR	**0.442**	0.397	0.336	<u>0.429</u>	0.425
Precision	**0.476**	0.437	0.380	<u>0.466</u>	0.462

The most optimal outcomes are indicated in **bold**, while the second-best results are underlined.

statistical significance of ITRPCA compared to other methods in terms of AUC, AUPR, and precision. The *p*-values were carefully adjusted using the Bonferroni correction to control for multiple testings. Table 4 shows the *p*-values obtained from the rank sum test and the Bonferroni correction. The results indicate that ITRPCA is significantly better than the other methods, except for MLMC (p-value < 0.05). It suggests that ITRPCA outperforms most of the compared methods in terms of AUC, AUPR, and precision. The significance of the comparison was carefully adjusted using the Bonferroni correction to control for multiple testing.

In addition, to demonstrate the computational efficiency of the compared methods, we have recorded the average amount of time taken by each fold. The 10-fold cross-validation is executed on a personal laptop, which is powered by an Intel Core i7 processor and comes with 16 GB RAM. Figure 2B shows the average running time for each of the 10 folds across all comparison methods. As shown in Figure 2B, the methods with an average running time of less than 10 seconds are HGBI, MSBMF, and ITRPCA. The average required time for MLMC and MBiRW is relatively long, which is approximately five times that of our method. Therefore, ITRPCA is a promising prediction method that shows both effective predictions and efficient computational performance.

4.4 Independent testing

To further demonstrate the performance of ITRPCA in real applications, we conduct two types of independent testing experiments. The gold standard dataset is used as the training set to train the models, and the set of associated pairs in the Cdataset excluding the training set is used as the testing set to evaluate the performance of the models. To be specific, we have collected a total of 57 drug–disease association pairs in the testing set. Figure 3 shows the ROC curves of all comparative methods in independent testing. As shown in Figure 3, ITRPCA has demonstrated clear superiority over other methods in this independent testing. Specifically, ITRPCA yields an AUC value of 0.943, while HGBI, MBiRW, BNNR, MSBMF, and MLMC yield AUC values of 0.873, 0.908, 0.882, 0.925, and 0.892, respectively. It is worth mentioning that the AUC value of ITRPCA is 5.717% higher than that of MLMC.

TABLE 6 Top 10 candidate indications for cisplatin, vincristine, doxorubicin, methotrexate, and cytarabine.

Drugs (DrugBank ID)	Top 10 candidate diseases (OMIM ID)
Cisplatin (DB00515)	Rhabdomyosarcoma 2 (268220); **lung cancer (211980);** lymphoblastic leukemia, acute, with lymphomatous features (247640); **diffuse gastric and lobular breast cancer syndrome (137215);** reticulum cell sarcoma (267730); leukemia, chronic lymphocytic, susceptibility to 2 (109543); **Wilms tumor 1 (194070); breast cancer (114480); colorectal cancer (114500);** thyroid cancer, and non-medullary, 2 (188470)
Vincristine (DB00541)	Leukemia, chronic lymphocytic (151400); mycosis fungoides (254400); myelofibrosis (254450); **breast cancer (114480); osteogenic sarcoma (259500); bladder cancer (109800); lung cancer (211980); Kaposi sarcoma, susceptibility to (148000); small cell cancer of the lung (182280);** and diffuse gastric and lobular breast cancer syndrome (137215)
Doxorubicin (DB00997)	Leukemia, chronic lymphocytic, susceptibility to, 2 (109543); reticulum cell sarcoma (267730); **esophageal cancer (133239);** small cell cancer of the lung (182280); testicular germ cell tumor (273300); **colorectal cancer (114500);** Dohle bodies and leukemia (223350); **prostate cancer (176807); renal cell Carcinoma, nonpapillary (144700); and hepatocellular carcinoma (114550)**
Methotrexate (DB00563)	**Lung cancer (211980);** Wilms tumor 1 (194070); **leukemia, chronic lymphocytic (151400); myeloma, multiple (254500); prostate cancer (176807);** renal cell carcinoma, nonpapillary (144700); neuroblastoma, susceptibility to, 1 (256700); thyroid cancer, nonmedullary, 2 (188470); myelofibrosis (254450); and moved to 619182 (175505)
Cytarabine (DB00987)	**Leukemia, chronic lymphocytic (151400);** mycosis fungoides (254400); **myelofibrosis (254450);** rhabdomyosarcoma 2 (268220); **myeloma multiple (254500);** colorectal cancer (114500); small cell cancer of the lung (182280); Kaposi sarcoma, susceptibility to (148000); testicular germ cell tumor (273300); and **breast cancer (114480)**

The predicted indications in **bold** have been confirmed by the CTD.

In addition, the other independent testing is conducted using all known associations in the gold standard dataset as training samples and unknown associations as candidate samples. The prediction scores of all candidate pairs are obtained by computational methods and ranked for each specific drug. We focus on how many of the top n candidate indications for each drug could be found and confirmed to have been used in clinical treatment in the CTD (released in February 2020) (Davis et al., 2019). Specifically, among all the drugs and diseases involved in the gold standard dataset, we have identified a total of 938 drug–disease associations that were subsequently validated in the CTD. As shown in Figure 4, the number of correctly predicted associations for 593 drugs is counted for the top 5 to top 30 candidate indications. It is evident that ITRPCA predicts the highest number of correct associations among all the methods for all drugs, followed by MSBMF and MLMC. Specifically, the number of validated associations from the top 5 to top 30 identified by ITRPCA is 163, 269, 348, 409, 456, and 492, respectively. In contrast, the corresponding numbers of identified associations by MLMC are all lower than those by ITRPCA, with a difference of 7, 30, 35, 44, 45, and 46, respectively.

4.5 Ablation experiment

To elucidate the individual impact of components within ITRPCA, we designed four ablation experiments: “w/o WKNN,” “only WKNN,” “ITRPCA-drug,” and “ITRPCA-disease.” To be specific, “w/o WKNN” implies the ITRPCA method without WKNN preprocessing for prediction. “only WKNN” represents using only the WKNN algorithm to infer the potential drug–disease associations, without the need for using our tensor RPCA model. “ITRPCA-drug” represents that only the drug tensor in ITRPCA was used to predict drug–disease associations, while “ITRPCA-disease” only uses the disease tensor in ITRPCA. To ensure a rigorous and unbiased comparison, the same prior similarity information and parameters as the ITRPCA model are employed in the aforementioned experiments.

Table 5 shows the AUC, AUPR, and precision results obtained from the 10-fold cross-validation of the comparative methods on the gold standard dataset. As anticipated, ITRPCA performs the best with AUC, AUPR, and precision values. This indicates that combining WKNN and TRPCA has a positive impact on predictive performance. In fact, the “w/o WKNN” model does not exhibit prominent results in predicting latent associations. It illustrates that WKNN preprocessing in ITRPCA can assist in the prediction. For the “only WKNN” model, relevant information was added based on drug and disease similarity. However, this addition also introduced more noise, leading to poor prediction performance. It serves as evidence from the opposite perspective that the effectiveness of TRPCA in noise reduction is significant. Furthermore, based on the prediction results of “ITRPCA-drug” and “ITRPCA-disease,” we find that the simultaneous utilization of tensor information from both drugs and diseases leads to better performance compared to using only one type of tensor information. It implies that the effective enhancement of prediction outcomes can be achieved through the integration of prior knowledge from drugs and diseases.

4.6 Case studies

To demonstrate the practical application of ITRPCA, we conducted case studies with the aim of uncovering novel applications for existing drugs. By utilizing all available drug–disease associations and multiple similarities in the gold standard dataset, we applied the ITRPCA method to predict the unexplored relationship between drugs and diseases. Based on the prediction results of ITRPCA, we generated all possible candidate indications for each drug and sorted them according to their obtained scores. In recent years, the development of drugs for tumors and leukemia has received widespread attention. Here, we selected four commonly used anti-tumor drugs (cisplatin, vincristine, doxorubicin, and methotrexate) and one anti-malignant hematologic drug (cytarabine) to search for evidence of their candidate indications in the CTD.

Table 6 shows the top 10 candidate indications predicted by the ITRPCA algorithm for the five drugs, with confirmed indications highlighted in bold. It was observed that each drug had 4–6 validated indications among the top 10 predictions. As an example, doxorubicin (DB00997), a broad-spectrum antitumor medication with antibiotic-like properties, was found to be effective in treating various types of cancer, including esophageal cancer, colon cancer, prostate cancer, renal cell carcinoma (nonpapillary), and hepatocellular carcinoma, as shown in Table 6. Additionally, chronic lymphocytic leukemia (susceptibility to, 2) and reticulum cell sarcoma were ranked first and second in the candidate indication list, respectively. However, their validity as indications has not been confirmed yet. These unconfirmed candidate indications hold potential as promising targets for further research.

5 Conclusion

In the study, we have proposed a novel computational method called ITRPCA for identifying drug-associated indications. ITRPCA can not only effectively exhibit robustness in isolating the low-rank tensor and noise information but also restrict predicted entry values of the low-rank tensor within a specific interval. The cross-validation and independent testing experiments have shown that ITRPCA is a highly effective prediction method. In particular, when compared to existing drug repositioning methods in independent testing, ITRPCA outperforms them in all measures, indicating a clear advantage. Additionally, case studies have confirmed ITRPCA's reliability in predicting new indications for known drugs. Therefore, we are confident that ITRPCA will serve as a valuable tool to successfully facilitate practical drug repositioning.

Data availability statement

Publicly available datasets were analyzed in this study. These data can be found at: https://github.com/YangPhD84/ITRPCA.

Author contributions

MY and BY wrote the manuscript. MY conducted the methodology and software development. BY performed the investigation and validation. GD and JW reviewed and edited the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by the National Natural Science Foundation of China (Grant Nos 62272309 and 61972423), the Key Project of Hunan Education Department (Grant No. 21A0478), the Natural Science Foundation of Hunan (Grant No. 2022JJ30549), and the Graduate Research Innovation Project of Shaoyang University (Grant No. CX2023SY053).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Ada, H., Scott, A. F., Amberger, J. S., Bocchini, C. A., and McKusick, V. A. (2005). Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* 33 (1), 514–517. doi:10.1093/nar/gki033

Ashburn, T. T., and Thor, K. B. (2004). Drug repositioning: identifying and developing new uses for existing drugs. *Nat. Rev. Drug Discov.* 3 (8), 673–683. doi:10.1038/nrd1468

Chong, C., and Sullivan, D. (2007). New uses for old drugs. *Nature* 448 (7154), 645–646. doi:10.1038/448645a

Davis, A., Grondin, C. J., Johnson, R. J., Sciaky, D., McMorran, R., Wiegers, J., et al. (2019). The comparative toxicogenomics database: update 2019. *Nucleic Acids Res.* 47 (D1), D948–D954. doi:10.1093/nar/gky868

Dickson, M., and Gagnon, J. P. (2009). The cost of new drug discovery and development. *Discov. Med.* 4 (22), 172–179.

Gao, C. Q., Zhou, Y. K., Xin, X. H., Min, H., and Du, P. F. (2022). DDA-SKF: predicting drug-disease associations using similarity kernel fusion. *Front. Pharmacol.* 12, 784171. doi:10.3389/fphar.2021.784171

Gao, Q., Zhang, P., Xia, W., Xie, D., Gao, X., and Tao, D. (2021). Enhanced tensor RPCA and its application. *IEEE Trans. Pattern Analysis Mach. Intell.* 43 (6), 2133–2140. doi:10.1109/TPAMI.2020.3017672

Gottlieb, A., Stein, G. Y., Ruppin, E., and Sharan, R. (2011). Predict: A method for inferring novel drug indications with application to personalized medicine. *Mol. Syst. Biol.* 7 (1), 496. doi:10.1038/msb.2011.26

Huang, F., Qiu, Y., Li, Q., Liu, S., and Ni, F. (2020). Predicting drug-disease associations via multi-task learning based on collective matrix factorization. *Front. Bioeng. Biotechnol.* 8, 218. doi:10.3389/fbioe.2020.00218

Jaccard, P. (1908). Nouvelles recheres sur la distribution florale. *Bull. Soc. Vaud. Sci. Nat.* 44, 223–270.

Jarada, T. N., Rokne, J. G., and Alhajj, R. (2021). SNF-NN: computational method to predict drug–disease interactions using similarity network fusion and neural networks. *BMC Bioinforma.* 22 (1), 28–20. doi:10.1186/s12859-020-03950-3

Jiang, H., and Huang, Y. (2022). An effective drug–disease associations prediction model based on graphic representation learning over multi-biomolecular network. *BMC Bioinforma.* 23, 9–17. doi:10.1186/s12859-021-04553-2

Kilmer, M. E., and Martin, C. D. (2011). Factorization strategies for third-order tensors. *Linear Algebra Its Appl.* 435 (3), 641–658. doi:10.1016/j.laa.2010.09.020

Lavecchia, A. (2015). Machine-learning approaches in drug discovery: methods and applications. *Drug Discov. Today* 20 (3), 318–331. doi:10.1016/j.drudis.2014.10.012

Liu, C., Wei, D., Xiang, J., Ren, F., Huang, L., Lang, J., et al. (2020). An improved anticancer drug-response prediction based on an ensemble method integrating matrix completion and ridge regression. *Mol. Therapy-Nucleic Acids* 21, 676–686. doi:10.1016/j.omtn.2020.07.003

Liu, G., Lin, Z., and Yu, Y. (2010). "Robust subspace segmentation by low-rank representation," in Proceedings of the 27th International Conference on Machine Learning (ICML-10), Madison, WI, United States, June 2010, 663–670.

Lu, C., Feng, J., Chen, Y., Liu, W., Lin, Z., and Yan, S. (2020). Tensor robust principal component analysis with a new tensor nuclear norm. *IEEE Trans. Pattern Analysis Mach. Intell.* 42 (4), 925–938. doi:10.1109/TPAMI.2019.2891760

Luo, H., Li, M., Yang, M., Wu, F. X., Li, Y., and Wang, J. (2021). Biomedical data and computational models for drug repositioning: A comprehensive review. *Briefings Bioinforma.* 22 (2), 1604–1619. doi:10.1093/bib/bbz176

Luo, H., Wang, J., Li, M., Luo, J., Peng, X., Wu, F. X., et al. (2016). Drug repositioning based on comprehensive similarity measures and Bi-Random walk algorithm. *Bioinformatics* 32 (17), 2664–2671. doi:10.1093/bioinformatics/btw228

Peng, Y., Ganesh, A., Wright, J., Xu, W., and Ma, Y. (2012). Rasl: robust alignment by sparse and low-rank decomposition for linearly correlated images. *IEEE Trans. Pattern Analysis Mach. Intell.* 34, 2233–2246. doi:10.1109/TPAMI.2011.282

Pham, D. H., Basarab, A., Zemmoura, I., Remenieras, J. P., and Kouame, D. (2021). Joint blind deconvolution and robust principal component analysis for blood flow estimation in medical ultrasound imaging. *IEEE Trans. Ultrasonics, Ferroelectr. Freq. Control* 68 (4), 969–978. doi:10.1109/TUFFC.2020.3027956

Qin, L., Wang, J., Wu, Z., Li, W., Liu, G., and Tang, Y. (2022). Drug repurposing for newly emerged diseases via network-based inference on a gene-disease-drug network. *Mol. Inf.* 41 (9), 2200001. doi:10.1002/minf.202200001

Resnik, P. (1995). "Using information content to evaluate semantic similarity in a taxonomy," in Proceedings of the 14th International Joint Conference on Artificial Intelligence, Montreal, Quebec, Canada, August 1995.

Steinbeck, C., Han, Y., Kuhn, S., Horlacher, O., Luttmann, E., and Willighagen, E. (2003). The Chemistry Development Kit (CDK): an open-source java library for chemo- and bioinformatics. *J. Chem. Inf. Comput. Sci.* 43 (2), 493–500. doi:10.1021/ci025584y

Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., et al. (2019). Applications of machine learning in drug discovery and development. *Nat. Rev. Drug Discov.* 18 (6), 463–477. doi:10.1038/s41573-019-0024-5

Van, D., Bruggeman, J., Vriend, G., Brunner, H. G., and Leunissen, J. A. M. (2006). A text-mining analysis of the human phenome. *Eur. J. Hum. Genet.* 14 (5), 535–542. doi:10.1038/sj.ejhg.5201585

Wang, H., Zhao, S., Zhao, J., and Feng, Z. (2021). A model for predicting drug–disease associations based on dense convolutional attention network. *Math. Biosci. Eng.* 18 (6), 7419–7439. doi:10.3934/mbe.2021367

Wang, J. Z., Du, Z., Payattakool, R., Yu, P. S., and Chen, C. F. (2007). A new method to measure the semantic similarity of GO terms. *Bioinformatics* 23 (10), 1274–1281. doi:10.1093/bioinformatics/btm087

Wang, S. J., Yan, W. J., Zhao, G., Fu, X., and Zhou, C. G. (2014). Micro-Expression Recognition Using Robust Principal Component Analysis and Local Spatiotemporal Directional Features. *European Conference on Computer Vision*. Berlin, Germany: Springer.

Wang, W., Yang, S., and Li, J. (2013). Drug target predictions based on heterogeneous graph inference. *Pac. Symposium Biocomput.* 18, 53–64. doi:10.1142/9789814447973_0006

Weininger, D. (1988). SMILES, a chemical language and information system. 1. introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* 28 (1), 31–36. doi:10.1021/ci00057a005

Wishart, D., Knox, C., Guo, A. C., Shrivastava, S., Hassanali, M., Stothard, P., et al. (2006). DrugBank: A comprehensive resource for *in silico* drug discovery and exploration. *Nucleic Acids Res.* 34, 668–672. doi:10.1093/nar/gkj067

Wright, J., Ganesh, A., Rao, S., Peng, Y., and Ma, Y. (2009). Robust principal component analysis: exact recovery of corrupted low-rank matrices by convex optimization. *Adv. Neural Inf. Process. Syst.* 22.

Xiao, Q., Luo, J., Liang, C., Cai, J., and Ding, P. (2018). A graph regularized non-negative matrix factorization method for identifying microRNA-disease associations. *Bioinformatics* 34 (2), 239–248. doi:10.1093/bioinformatics/btx545

Xie, Y., Gu, S., Liu, Y., Zuo, W., Zhang, W., and Zhang, L. (2016). Weighted Schatten p-Norm Minimization for Image Denoising and Background Subtraction. *IEEE Trans. Image Process.* 25 (10), 4842–4857. doi:10.1109/tip.2016.2599290

Xuan, P., Ye, Y., Zhang, T., Zhao, L., and Sun, C. (2019). Convolutional neural network and bidirectional long short-term memory-based method for predicting drug-disease associations. *Cells* 8 (7), 705. doi:10.3390/cells8070705

Yan, Y., Yang, M., Zhao, H., Duan, G., Peng, X., and Wang, J. (2022). Drug repositioning based on multi-view learning with matrix completion. *Briefings Bioinforma.* 23 (3), bbac054. doi:10.1093/bib/bbac054

Yang, J., He, S., Zhang, Z., and Bo, X. (2021a). NegStacking: drug-Target interaction prediction based on ensemble learning and logistic regression. *IEEE/ACM Trans. Comput. Biol. Bioinforma.* 18 (6), 2624–2634. doi:10.1109/TCBB.2020.2968025

Yang, M., Luo, H., Li, Y., and Wang, J. (2019). Drug repositioning based on bounded nuclear norm regularization. *Bioinformatics* 35 (14), i455–i463. doi:10.1093/bioinformatics/btz331

Yang, M., Wu, G., Zhao, Q., Li, Y., and Wang, J. (2021b). Computational drug repositioning based on multi-similarities bilinear matrix factorization. *Briefings Bioinforma.* 22 (4), bbaa267. doi:10.1093/bib/bbaa267

Yu, Z., Huang, F., Zhao, X., Xiao, W., and Zhang, W. (2021). Predicting drug–disease associations through layer attention graph convolutional network. *Briefings Bioinforma.* 22 (4), bbaa243. doi:10.1093/bib/bbaa243

Zhang, L., and Peng, Z. (2019). Infrared small target detection based on partial sum of the tensor nuclear norm. *Remote Sens.* 11 (4), 382. doi:10.3390/rs11040382

Zhao, B. W., Hu, L., You, Z. H., Wang, L., and Su, X. R. (2022). Hingrl: predicting drug–disease associations with graph representation learning on heterogeneous information networks. *Briefings Bioinformatic* 23 (1), bbab515. doi:10.1093/bib/bbab515